Koartikulációs modellek a magyar nyelvű gépi beszédfelismerésben

Mihajlik Péter

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
mihajlik@tmit.bme.hu

Kivonat: A koartikulációs jelenségek két elvi csoportjának, a fonológiai és a fonetikai koartikuláció modellezésének kérdéseit, megoldásait vizsgáljuk magyar nyelvű statisztikai-alapú gépi beszédfelismerés esetén. A koartikulációs modellek beszédfelismerési hálózatba integrálásának érdekében bevezetjük a súlyozott véges állapotú átalakító (WFST) alapú felismerési hálózatépítést. Bemutatjuk az alkalmazott explicit és implicit koartikulációs modelleket, melyeket az általunk elérhető legnagyobb magyar nyelvű – részben publikus – telefonbeszéd-adatbázisok segítségével értékelünk ki. Az eredmények meggyőzően mutatják, mely típusú koartikuláció modellezésére érdemes nagyobb hangsúlyt fektetni a folyamatos beszédfelismerési pontosság jelentős növelésének érdekében.

1 Bevezetés

A koartikuláció – az egymást követő hangok egymásra hatása, „együtt ejtése” – a beszéd alapvető jellegzetessége. Nem különbözik e tekintetben a magyar nyelv más nyelvektől, amit az is jelez, hogy a korszerű magyar nyelvű gépi beszédelőállítási módszerek mindegyike elemi vagy magasabb szinten explicit hangkapcsolati modelleket használ (diádok, triádok, stb.) [5].

A magyar nyelvű gépi beszédfelismerés területén ugyanakkor – a nemzetközi trendekkel ellentétben – a különböző koartikulációs jelenségek explicit modellezése nem jellemző. Tipikus a környezetfüggetlen modellek használata, melyeknél a fonéma – beszédhang szintek szétválasztása fizikailag nem történik meg. A kutatócsoportunkhoz kötődő publikációkon felül nem ismerünk kísérleteket a magyar nyelvű koartikulációs jelenségeket explicit módon kezelő beszédhang-modellezéssel kapcsolatban.

A következőkben rövid áttekintést kívánunk nyújtani a magyar nyelv koartikulációs jelenségeinek modellezésével kapcsolatos nehézségekről, jellegzetességekről, és megoldásairól. Az egyes megközelítések hatását a beszédfelismerés pontosságára a legnagyobb elérhető magyar nyelvű beszédadatbázisokon különféle konfigurációkban és felismerési feladatokban mértük. Az eredmények megbízhatóságát szignifikancia-vizsgálattal ellenőriztük.

2 A koartikulációs jelenségek osztályozása

2.1 Fonológiai koartikulációs jelenségek

A modern nyelvtudomány a "kiejtési szabályok" néven összegyűjtött hasonulási, összeolvadási, stb. jelenségeket fonológiai koartikulációs jelenségeknek hívja. Ezek főbb ismérve, hogy egy vagy több beszédhang fonémaértéke megváltozik a kiejtés során (pl. *azt → a sz t*). A megváltozás lehet összetettebb jelenség, beleértve a kiesést vagy betoldást is (pl. *értsd → é r dzs d, tea → t e j a*). Külön említendők a szóhatárokon fellépő fonológiai változások (pl. *értds te → é r dzs d _ t e* vagy *é r cs t e*), melyek attól is függhetnek, hogy tart-e szünetet a beszélő a két szó között vagy sem, illetve, természetesen attól is, hogy milyen hanggal kezdődik a következő szó.

A fonológiai koartikulációs jelenségek egy lehetséges csoportosítása a következő:

- Zöngésségi (részleges és teljes) hasonulások: *adta → a tt a, lékbe → l é g b e*
- Képzés helye szerinti (részleges és teljes) hasonulások:
  azonban → a z o m b a n, önmaga → ö mm a g a
- Mássalhangzó-rövidülések: *állt → á l t*
- Összeolvadások: *látja → l á tty a, utca → u cc a, kétség → k é ccs é g*
- Egyéb kiesések, betoldások:
  parasztkolbász → p a r a sz k o l b á sz, tea → t e j a

2.2 Fonetikai koartikulációs jelenségek

A fonetikai koartikuláció a beszéd nagyon fontos jelensége. Lényege, hogy a beszélőszervek tehetetlenégének, folyamatos mozgásának következtében a hangátmenetek nagy része is folyamatos, így a beszédhangok jelentős része az önmagában való ejtéshez képest megváltozik. A fonetikai koartikuláció segít például a felpattanó zárhangok felismerésénél, ahol a környező magánhangzók formánsátmenetei engednek következtetni a zárhang identitására.

A fonetikai koartikuláció természetesen szóhatárokon is felléphet, illetve a beszédszünet is hatással lehet a környező hangokra.

3 A koartikulációs jelenségek modellezése

3.1 A WFST keretrendszer

Az előzőekben leírt koartikulációs jelenségek modellezése különösen a folyamatos beszédfelismerés esetén jelent nagy kihívást, hiszen attól függően lép fel egyik vagy másik jelenség, hogy az adott szó után melyik másik következik. Egyedi specializált megoldások helyett a súlyozott véges állapotú átalakítókkal (WFST – Weighted

Finite-State Transducers) történő tudásforrás reprezentációt és integrációt választottuk, mely általános, matematikailag is letisztult keretet biztosít a feladathoz.

A súlyozott véges állapotú átalakítók formálisan a *félgyűrűk* felett értelmezett matematikai objektumokként definiálhatók [4]. Praktikusan a véges állapotú gépek olyan általánosításának tekinthetők, melyek egy adott bejövő szimbólumsorozatnak nem csak az elfogadásáról vagy elvetéséről dönthetnek, hanem képesen súlyt és kimeneti szimbólumsorozatot is rendelni hozzájuk.

A WFST-keretrendszerben a beszédfelismerés során használt tudásforrásokat – úgymint nyelvi modell, kiejtési szótár, beszédhangmodellek, stb. – először súlyozott véges állapotú átalakító formára kell hozni, majd ezeket standard WFST műveletekkel lehet egybe komponálni és optimalizálni. Két tudásforrás ötvözésére a kompozíció (jelölés: o), az egyes tudásforrások optimalizálására pedig a determinizáció és minimalizáció használható (jelölés: det, min) [4].

A felismerési hálózat összeállításának szemléltetése környezetfüggetlen beszédhang-modellezés esetén:

$$\textit{Felismerési hálózat} = \min(\mathrm{H} \circ \mathrm{L} \circ \mathrm{G}), \qquad (\mathbf{1})$$

ahol G: a nyelvi modell
L: a kiejtési modell
H: a fonémák leképezése elemi akusztikus modellekre

A felismerési hálózat ilyenkor HMM (rejtett Markov-modell) állapot szimbólumsorozatot képez le szósorozatra a nyelvi, kiejtési és egyéb súlyoknak megfelelően, így közvetlenül használható a „hagyományos” Viterbi-féle HMM dekódolási algoritmus a felismerési eredmények valós idejű meghatározásához.

A keretrendszer óriási előnye a flexibilitás. Bármilyen kiejtési alternatívákat valószínűségekkel ellátó, vagy a legegyszerűbb fonológiai kiejtési modell integrálható, csakúgy mint a bigram helyett trigram vagy 4-gram nyelvi modell a rendszer bárminemű megbontása nélkül. Hátránya, hogy az inkrementális hálózatépítés (új szó hozzáadása) alapesetben a teljes felismerési hálózat újraépítését teszi szükségessé, mely jelentős számítási igénnyel járhat.

3.2 Fonológiai koartikulációs modellek

A hasonulási, egybeolvadási stb. fonológiai koartikulációs szabályok speciális esetei a környezetfüggő újraíró szabályoknak. Ezek WFST implementációjáról részletesen szól a [4], de az egyedi implementálás is járható út.

Az egyes zöngésségi, képzés helye szerinti hasonulások és opcionális összeolvadások súlyozott véges állapotú átalakítóinak kompozíciójával kaphatjuk meg az általános fonológiai koartikulációs modell WFST reprezentációját

A kísérletekben a [3]-ban bemutatott hierarchikus fonológiai koartikulációs szabályrendszer WFST megfelelőjét használtuk, melyet a következő oldalon részletezett módon állítottunk össze elemi szabálytípusoknak megfelelő véges átalakítókból.

P_1: Zöngésségi hasonulás /kötelező/

P_2: Összeolvadás + Rövidülés /kötelező/

P_3: Képzés helye, módja szerinti részleges hasonulások /opcionális/

P_4: Képzés helye, módja szerinti teljes hasonulások /opcionális/

Az fonológiai koartikulációs modell, P, [3] után az alábbi kompozíció-sorozattal adódik:

$$P = P2 \text{ o } P4 \text{ o } P3 \text{ o } P2 \text{ o } P1 \qquad (\mathbf{2})$$

Ez a modell a 2.1-ben említett fonológiai koartikulációs jelenségek közül mindegyiket explicit módon, *szóhatárokon átívelve* (is) kezeli. Kivételt csak az "egyéb kiesések, betoldások" képeznek, mert ezek esetlegesek, ritkák és automatizáltan nem állíthatók elő. Megjegyezzük, hogy a szóhatárokon átívelő koartikulációt csak akkor tesszük lehetővé, ha a két szó közé nem esik szünet a kiejtés során.

3.3 Fonetikai koartikulációs modellek

A fonetikai koartikuláció explicit modellezésére a környezetfüggő beszédhangmodelleket, ezen belül is a szóhatárokon átívelő („cross-word"), mindkét oldalon 1 hang távolságig környezetfüggő (trifón) modelleket választottuk. A trifón modelleket 3 állapotú, „left-to-right" struktúrájú rejtett Markov-modellek (HMM-ek) képviselik. Így a koartikulációt beszédhangonként a fonetikus környezettől függő kialakítású 3 kiejtési fázissal modellezzük.

Az általánosított trifón modellek állapotait fonémánként és állapotonkénti ML (Maximum Likelihood) fonetikus döntési fákkal csoportosítottuk [7]. Mivel az eljárás magyar nyelv esetén még nem bevett, ugyanakkor mind elméleti, mind gyakorlati szempontból fontos, a következőkben röviden vázoljuk.

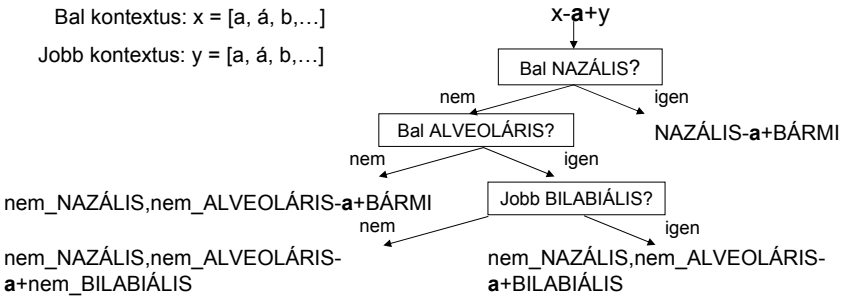
A módszer lényege, hogy az adott fonéma adott állapotához tartozó általánosított trifón állapotokat a kezdeti egy csoportból lépésről-lépésre úgy osztja további csoportokra, hogy a felhasználó által definiált fonetikus környezetre utaló kérdéseket sorban felteszi, és végül azt választja, mely ML értelemben a legjobb szeparációt jelenti. Az eljárás akkor ér véget, amikor egy csoportra már nem jut elég tanítóminta, vagy az új csoport kettéosztás már nem hoz érdemi hasonlósági mérték növekedést a tanítóadatbázison. Végeredményben egy döntési fa áll elő minden fonéma bal, középső és jobb állapotára (esetünkben összesen 3x39 döntési fa), melynek levelei reprezentálják a trifón állapot csoportokat.

Az eljárást a következő példával szemléltetjük.

Legyenek a bal és jobbkörnyezetekre utaló kérdések az alábbi módon definiálva:

NAZÁLIS: m, n, …
ALVEOLÁRIS: d, t, n, …
VELÁRIS: g, k, …
BILABIÁLIS: p, b, m…

Egy lehetséges döntési fa az „a” hang bal szélső állapotára:



1. ábra. A fonetikai döntési fa alapú trifón állapot csoportosítás szemléltetése.

Szemléltető példaként tekintsük a „pamacs” szó első és második „a” hangjának elemi akusztikus modellekre (HMM állapotokra) történő leképezését. Az első „a” általánosított trifónja a „p-a+m” a másodiké a „m-a+cs”: ezeket a fenti döntési fán kiértékelve kapjuk meg az adott – esetben a bal – állapot elemi akusztikus modelljét, mely a csoporthoz tartozó mintákkal lett tanítva. A „p-a+m” bal szélső állapota a fenti döntési fa alapján a „nem_NAZÁLIS,nem_ALVEOLÁRIS-**a**+BÁRMI” trifón állapot csoportba, míg az „m-a+cs” bal állapota pedig „NAZÁLIS-**a**+BÁRMI” csoportba kerül.

Az ML döntési fa-alapú trifón állapotcsoportosítás jó tulajdonsága, hogy a csoportok avagy az elemi akusztikus modellek száma két – a fa-építés leállításánál említett – küszöbérték segítségével széles határok közt állítható. Továbbá, hogy a tanító-adatbázis méretéhez jól alkalmazkodik, kisebb adatbázis esetén kevesebb, nagyobb adatbázisnál nagyobb számú elemi akusztikus modellre képez le azonos küszöbértékek esetén is. A megközelítés hátránya, hogy döntési fa építéshez igényel egy általánosított trifón szintű akusztikai modell tanítást is a tanító beszédadatbázison.

A fonetikai koartikulációs modell WFST-formátumra hozása két lépésben történhet. Az első lépés a fonémasorozat - általánosított trifónsorozat leképezés, melyet a CD véges átalakító végez. Súlyokra itt nincs szükség, mivel a leképezés egyértelmű. A CD átalakító képzésének kifejtésére itt nem vállalkozhatunk, az a [4]-ban megtalálható. A következő lépés az általánosított trifónok elemi akusztikus modellekre (HMM állapotokra) való leképzése, melyet a H_{tri} véges átalakító hajt végre. Ehhez az összes lehetséges általánosított trifón megfelelő döntési fán való kiértékelése szükséges, melynek eredménye egy trifón kiejtési táblázatba foglalható, amely már triviálisan alakítható véges átalakítóvá.

A fonéma-HMM állapot leképezés a környezetfüggő beszédhangmodelleknél tehát a H_{tri} o CD kompozícióval adódik, ahol a H_{tri} kialakítása tanító-adatbázis függő.

4 A koartikulációs modellek kiértékelése

A koartikulációs modellek kiértékelésénél természetes választás a beszédfelismerési tesztekkel történő minősítés. Ilyenkor fontos, hogy a felismerési feladat elég általános

legyen, valamint, hogy a teszt (és tanító) adatok elég változatosak, nagyszámúak és reprezentatívak legyenek. Továbbá a tanító és teszt adatok függetlenségének biztosítása is kívánatos, vagy legalábbis ennek kézbentartása.

A felismerési tesztek eredményei azonban önmagukban nemigen használhatók, ezért minden kísérletnél összehasonlításokat végeztünk. Általában az előző fejezetben tárgyalt explicit koartikulációs modelleket hasonlítottuk össze az implicit modellekkel. Fonológiai koartikulációnál az implicit modellt az jelentette, amikor a fonológiai szinten tanításnál sem vettük figyelembe a koartikulációs jelenségeket. A fonetikai koartikulációnál pedig az implicit modell a monofón, azaz a környezet független beszédhang-modell volt.

A következőkben röviden összefoglaljuk a kísérleti körülményeket, majd az elvégzett vizsgálatok lépéseit és eredményeit.

4.1 Kísérleti körülmények

Beszédadatbázisok:

A tanító- és tesztelő-adatbázisokat a legnagyobb magyar telefonos beszédadatbázisok, az MTBA, a Besztel, a SpeechDat és a Tesztel összességéből alakítottuk ki [6]. Ezek az adatbázisok elsősorban olvasott beszédet, valamint kisebb arányban spontán bemondásokat is tartalmaznak. Az első három adatbázis lényegében ugyanarra a szövegkorpuszra épül, és mindegyiknek az általunk elérhető része 500 beszélőtől tartalmaz hanganyagot. A Tesztel adatbázis 100 beszélős, és jellegzetessége, hogy szándékosan nagy és természetes háttérzajban felvett bemondásokat tartalmaz. Az adatbázisokban a vonalas és mobil telefonos felvételek összességében körülbelül ugyanolyan számban képviseltetik magukat.

Tanítóhalmazok:

Tanítás céljára az MTBA, Besztel, és a SpeechDat adatbázis 500-400-450 beszélőjének azon felvételeit jelöltük ki, melyek nem „o”, és „z” jelzésűek, azaz nem tartalmaznak tulajdonneveket és bizonyos típusú mondatokat. A SpeechDat esetén csak egy szűkebb halmazt, a fonetikailag változatos szavakat és mondatokat (kivéve a „z” jelzésűeket) használtuk.

A teljes tanítóhalmaz mellett annak bizonyos részhalmazait is képeztük, hogy a különféle koartikulációs modellezési eljárások tanító adatbázisméret-függését is vizsgálhassuk.

Sem a tanítóhalmazokban, sem a későbbi teszthalmazokban nem végeztünk szűrést az annotációnál zajosnak minősített felvételekre. Kizárólag azokat a felvételeket hagytuk ki, melyeknek az eleje vagy vége az annotáció szerint nem került rögzítésre.

A tanítóhalmazok jelölése és tartalma:

- **M**: Az MTBA fonetikailag változatos mondatai és szavai, 500 beszélő, 6000 felvétel
- **MM**: Az MTBA összes tanítófelvétele, 500 beszélő, 19000 felvétel.
- **MM_BS**: Az MTBA és a Besztel összes tanítófelvétele, 900 beszélő, 39000 felvétel.
- **MM_BS_SD**: Az MTBA, a Besztel és a SpeechDat tanítófelvételei, 1350 beszélő, 44000 felvétel.

A felismerési feladat:

Az általános tapasztalat szerint a beszélőfüggetlen folyamatos beszédfelismerés támasztja a legnagyobb igényeket a kiejtési – koartikulációs modellekkel szemben. Ezért olyan *általános* folyamatos beszédfelismerési feladatot próbáltunk definiálni, ami a rendelkezésre álló adatbázisokkal megvalósítható. Természetesen adódott, hogy az adatbázisok azon mondatait tartalmazó bemondásokat ismertessük fel, melyek nem szerepelnek a tanító halmazokban. A beszélőfüggetlenség követelménye miatt azon felvételeket is ki kellett zárnunk, melyeknek a beszélőjét felhasználtuk a tanítás során.

Teszthalmazok:

A teszthalmazokat tehát úgy állítottuk össze, ne legyen átfedés a tanítóhalmazban szereplő beszélőkkel. Így a tanításnál fel nem használt 170 beszélőtől (Besztel 100, SpeechDat 50, TeszTel 20) kerültek felvételek a teszthalmazokba. Összesen 2385 felvételt kaptunk, melyeket a tanító-adatbázishoz való illeszkedés mértéke szerint két halmazra bontottunk.

A folyamatos beszédfelismerésnél fonológiai és nyelvi illeszkedésről is beszélhetünk. Az egyik halmazba azokat a mondatokat válogattuk, amelyeknek *szöveges tartalma* egyezett az akusztikus modelltanításnál használt mondatokéval (fonológiai illeszkedés), valamint amelyeknek szöveges tartalma a nyelvi modell tanításakor is felhasználásra került (nyelvi illeszkedés), ez az „illeszkedő” halmaz. A másik, „nem illeszkedő” teszthalmazba azok a felvételek kerültek, melyek szövegtartalma sem az akusztikus, sem a nyelvi modell tanításkor nem lett felhasználva. Egyéb halmazt nem vizsgáltunk.

A teszthalmazok jelölése és tartalma:

- **I**: Nyelvi és fonológiai szempontból a tanításhoz illeszkedő mondatok, 170 beszélő, 1973 felvétel: „s” jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s1” és „s2” jelzésű mondatok a TeszTel-ből.
- **N**: Sem nyelvi és sem fonológiai szempontból a tanításhoz nem illeszkedő mondatok, 170 beszélő, 412 felvétel: „z” jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s3” jelzésű mondatok a TeszTel-ből.

Beszédfelismerési paraméterek, beállítások:

Lényegkiemelés: Lényegkiemelési paraméterekként a bemenő beszédjelből MFCC (Mel Frequency Cepstral Coefficients) 12 dimenziós vektorokat képeztünk, melyekhez *logE* (keretenkénti logaritmikus energia) paramétert is csatoltunk, majd dinamikus Delta és Delta-Delta értékeket számítottunk. A statikus energiát végül kicsatolva összesen 38 dimenziós jellemzővektorokat kaptunk. Mind a tanítás, mind a tesztelés során alkalmaztuk a vak csatornakiegyenlítés (Blind Equalization) módszerét [2].

Elemi akusztikus modellek: Az atomi modellek rejtett Markov-modell állapotok voltak rögzített hurok és továbblépési valószínűségekkel. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket használtunk.

Fonetikai koartikulációs modellek (H_{mono} és H_{tri} o CD): Mind a monofón mind a trifón modelleknél a beszédhangokat 3 elemi akusztikus modellre képeztük le, az előbbi esetben a környezettől függetlenül az utóbbi esetben az ML döntési fa alapján a fonetikus környezettől függően. A döntési fákat - és így a H_{tri} leképezést - tanítóhalmazonként és fonológiai modellenként újraépítettük.

Fonológiai koartikulációs modell (P): A 3.2-ben ismertetett módon állítottuk össze a „kiejtési szabályok" néven közismert fonológiai koartikulációs jelenségek túlnyomó részét modellező véges állapotú átalakítót.

Lexikai modell (L): A kiejtési modellek nyers, fonológiai koartikulációkat nem tartalmazó fonemikus átiratait automatikusan állítottuk elő. Allofónikus változatokat nem jelöltünk, továbbá a hosszú és rövid mássalhangzókat sem különböztettük meg. Így – a szünetmodelleket nem számítva – összesen 39 fonológiai kategóriát használtunk. A szünetmodell háromállapotú környezetfüggetlen modell volt.

Az alkalmazott 5561 elemű szótár az összes előforduló szót tartalmazta (beleértve az illeszkedő és a nem illeszkedő teszthalmaz szavait), így szótáron kívüli elemek kezelésére nem volt szükség.

Nyelvi modell (G): A folyamatos felismerésnél szó-trigram nyelvi modelleket alkalmaztunk Katz-féle visszametszéssel és Good-Turing valószínűség-újraelosztással [1]. A tanítószöveg az illeszkedő tesztmondatok szövege alapján készült úgy, hogy minden különböző mondatot csak *egyszer* szerepeltettünk. Így az illeszkedő mondatokon PP=40-es perplexitást, a nem illeszkedő tesztmondatokon PP=6230-as (nagyon magas, azaz igen kedvezőtlen) perplexitás értéket kaptunk.

4.2 A fonológiai koartikulációs modellek kiértékelése manuálisan, fonémaszinten szegmentált tanító-adatbázis mellett

Első lépésként a fonológiai koartikuláció modellezés vizsgálatát tűztük ki célul adott, beszédhang-szinten *kézzel szegmentált és ellenőrzött* tanítóadatbázis-feldolgozás mellett. Erre egyedül az M-jelű tanítóhalmaz volt alkalmas (MTBA, fonetikailag változatos szavak, mondatok).

Explicit fonetikai koartikulációs modellezés – azaz trifón modellek – mellett végeztük az összehasonlítást, mert egyéb vizsgálataink szerint (lásd a 4.4 pontot) ez jelentette a nem vizsgált paraméterek optimális beállítását.

Az alábbi két felismerési hálózattal végeztünk kísérleteket:

- H_{tri} o CD o L o G – nincs fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit fonológiai koartikuláció-modellezés

Mivel a beszédhang-modelleket kézzel ellenőrzött – tehát a fonológiai koartikulációkat jelölő – fonetikus szegmentáció mellett tanítottuk, azok nem modellezték még implicite sem a fonológiai koartikulációs jelenségeket. Így a P modell alkalmazásától szignifikáns javulást vártunk. Az eredményeket az 1. és 2. táblázat mutatja.

1. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok folyamatos beszédfelismerési eredményei manuálisan szegmentált tanító-adatbázis mellett.[75]

a) I (PP = 40)	FA	FP
H_{tri} o CD o L o G	93.05	*91.40*
H_{tri} o CD o P o L o G	93.99	*92.57*
ΔH	-13.6	*-13.6*

b) N (PP = 6230)	FA	FP
H_{tri} o CD o L o G	60.84	*49.45*
H_{tri} o CD o P o L o G	62.02	*51.09*
ΔH	-6.1	*-3.2*

Az illeszkedő teszthalmaz esetén kétszámjegyű relatív hibacsökkenés (ΔH) figyelhető meg, ugyanakkor a nem illeszkedő halmaz esetén a javulás szerényebb. Az eredmények szignifikanciáját *2 mintás Z-próba* segítségével ellenőriztük. 5% szignifikanciaszint mellett (95% konfidencia szint) az illeszkedő teszthalmaz esetén valóban szignifikáns javulást tapasztaltunk, míg a nem illeszkedőnél nem.

Az I-hez képest az N teszthalmazon – ugyanazon felismerési feladatban – mért sokkal gyengébb felismerési eredményeket a vonatkozó igen magas nyelvi modell perplexitás (PP) magyarázza.

4.3 A fonológiai koartikulációs modellek kiértékelése következetes tanító-adatbázis feldolgozás mellett

Az előző vizsgálatnál a referencia rendszerben egyáltalán nem modelleztük a fonológiai koartikulációs jelenségeket, mégis csak az egyik teszthalmaznál kaptunk szignifikáns javulást az explicit modell alkalmazásával. Ezért felmerült a kérdés, hogy következetes gépi szegmentációt alkalmazva és nagyobb tanító-adatbázisokat használva is tapasztalható-e érdemi felismerési hiba csökkenés a P véges átalakítónak köszönhetően.

A következő gépi fonetikus szegmentációs módszert dolgoztunk ki a következetes fonológiai modellezés érdekében. A legnagyobb tanítóhalmazra (MM_BS_SD) képeztük a lineáris Gtr „nyelvi modellt”, majd előállítottuk a *tanítóadatokra* vonatkozó felismerési hálózatokat:

- H_{tri} o CD o L o Gtr – *implicit* fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o Gtr – explicit fonológiai koartikuláció-modellezés

Kezdeti beszédhangmodelleket tanítottunk be az M tanító halmaz manuális szegmentációja alapján. Ezekkel kényszerített felismerést („forced alignment”) végezve megkaptuk a fonológiai koartikulációt implicit valamint explicit módon tartalmazó gépi fonetikus szegmentációkat.

A különböző tanítóhalmazok és felismerési hálózatok esetén mindig a megfelelő tanítású beszédhangmodelleket alkalmaztuk. Összesen tehát 4x2 akusztikus modell halmazt vizsgáltunk 2 felismerési hálózattal.

---

[75] *Magyarázat a táblázatokhoz:*
FA: felismerési arány [%]. FP: felismerési pontosság [%]. ΔH: a hiba relatív megváltozása [%]. A definíciók részletezését lásd a [7]-ben.

A felismerési hálózatok a 4.2-ben vizsgáltakkal azonosak voltak:

- H_{tri} o CD o L o G – *implicit* fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit fonológiai koartikuláció-modellezés

Fontos megjegyezni, hogy a következetes tanítás és tesztelés miatt a P modell kihagyása már nem jelenti azt, hogy a fonológiai koartikulációt egyáltalán nem, hanem, hogy implicite, vagyis alacsonyabb, beszédhang szinten modellezzük.

2. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok folyamatos beszédfelismerési eredményei következetes gépi adatbázis-feldolgozás mellett.

a) I (PP = 40)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{tri} o CD o L o G	94.13	*92.54*	93.82	*91.97*	94.22	*92.55*	94.47	*92.93*
H_{tri} o CD o P o L o G	94.24	*92.69*	93.41	*91.66*	94.14	*92.54*	94.78	*93.05*
ΔH	-1.9	*-2.0*	+6.6	*+3.9*	+1.4	*+0.1*	-5.6	*-1.7*

b) N (PP = 6230)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{tri} o CD o L o G	61.95	*48.16*	61.27	*47.66*	64.24	*52.34*	64.42	*52.02*
H_{tri} o CD o P o L o G	62.34	*50.14*	61.24	*48.20*	64.09	*51.91*	65.13	*53.20*
ΔH	-1.0	*-3.8*	+0.07	*-1.0*	+0.4	*+0.9*	-2.0	*-2.5*

Ahogy a 2 a) és b) táblázatok mutatják, az implicit és explicit fonológiai koartikulációs modellek beszédfelismerési eredményei között a különbség minimális. A sziginfikancia-vizsgálatok egyetlen esetben sem mutattak ki érdemi különbséget, sőt a hiba nem is csökkent minden esetben.

Észrevehető, hogy a kézi helyett gépi fonetikus szegmentáció az M halmaz esetében nem rontott, hanem még javított is az eredményeken. Gyakorlatilag tehát következetes gépi tanitóadatbázis-feldolgozás mellett ugyanolyan jó eredmények érhetők el fonológiai modell *nélkül* is, mint a manuálisan szegmentált adatbázissal *és* explicit, szóhatárokon átívelő hasonulási modellel.

Nem várt tapasztalat volt ugyanakkor, hogy a tanítóadatbázis méretének növelése alig javított a felismerési eredményeken annak ellenére, hogy minden tanítási konfigurációban újraépítettük a trifón állapotcsoportosítást végző ML döntési fákat. Itt a nagyobb tanítóhalmazoknál a tesztfelvételekhez képesti nagyobb fonológiai illesztetlenség, illetve az adatbázisok gyakorlatilag közös szövegkorpuszra épülése lehetnek a mögöttes okok.

4.4 A fonetikai koartikulációs modellek kiértékelése

A *fonetikai* koartikulációs modellek kiértékelésekor a fenti tapasztalatok alapján az implicit fonológiai koartikuláció kezelést választottuk. Ez a gyakorlatban azt jelentette, hogy a tanító-adatbázis gépi szegmentálásához csakúgy, mint a tesztekhez a P-modell alkalmazása nélkül készítettük a felismerési hálózatokat. Vagyis következetes gépi tanítóadatbázis-feldolgozást alkalmaztunk.

A vizsgált felismerési hálózatok tehát a következők voltak:

- H_{mono} o CD o L o G – implicit *fonetikai* koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit *fonetikai* koartikuláció-modellezés

A monofón, vagy környezetfüggetlen beszédhangmodellek a folyamatos megfigyelési sűrűségfüggvényeik révén implicite modellezik a fonetikai koartikulációt, hiszen több fonetikus környezet mellett történik a tanításuk. A kérdés az, hogy ez az implicit modell hasonlóan viselkedik-e, mint a *fonológiai* koartikulációnál az implicit modell.

A különféle adatbázis-méretek és teszthalmazok melletti eredményeket mutatja a következő táblázat.

3. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok felismerési eredményei implicit és explicit *fonetikai* koartikulációs modellek mellett.

a) I (PP = 40)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{mono} o L o G	85.34	*84.34*	80.19	*78.60*	79.87	*78.54*	80.74	*79.52*
H_{tri} o CD o L o G	94.13	*92.54*	93.82	*91.97*	94.22	*92.55*	94.47	*92.93*
ΔH	-60	*-52*	-69	*-62*	-71	*-65*	-71	*-65*

b) N (PP = 6230)	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	FA	FP	FA	FP	FA	FP	FA	FP
H_{mono} o L o G	29.22	*23.65*	25.40	*19.36*	24.72	*18.54*	25.58	*20.54*
H_{tri} o CD o L o G	62.34	*50.14*	61.24	*48.20*	64.09	*51.91*	65.13	*53.20*
ΔH	-50	*-38*	-52	*-40*	-53	*-42*	-54	*-40*

Mint láthatjuk az implicit és explicit fonetikai koartikulációs modellezés közti különbség drámai, és természetesen minden esetben szignifikáns a hiba csökkenése a „legszigorúbb", 0.01%-os szignifikancia küszöb mellett is.

Felhívjuk a figyelmet a nem illeszkedő teszthalmaz abszolút felismerési eredményeire. Itt nem a felismerési hibák, hanem a *felismerési arányok és pontosságok* között figyelhető meg 2 – 3-szoros különbség.

Az is észrevehető, hogy a környezetfüggetlen beszédhang-modelleknél az adatbázisméret növelése szinte csak rontott a felismerési eredményeken.

Terjedelmi korlátok miatt nem tudjuk részletesen közölni a monofón modellek esetén a P fonológiai koartikulációs modell alkalmazása mellett mért eredményeket. Röviden összefoglalva, abban az esetben relatíve nagyobb mértékben javultak a felismerési mutatók, mint a trifón esetben, azonban az abszolút felismerési arányok továbbra is messze leszakadva követik csak az explicit fonetikai koartikulációs modellek eredményeit.

A kísérletekben a beszédfelismerés számításigénye a trifón modellek esetén kb. 1.5x-ös volt a monofón modellekhez képest, azonban hatékony felismerési hálózat-optimalizációval ezt később 0.8-as, tehát a monofón modellekhez képes *gyorsabb* szintre tudtuk vinni azonos felismerési hiba mellett. Az abszolút számítási igény P4 3GHz-es számítógépen tipikusan valós idő alatt mozgott.

5 Összefoglalás

A koartikuláció két elvi csoportjának, a *fonológiai* és a *fonetikai* koartikuláció modellezésének kérdéseiről, megoldásairól értekeztünk magyar nyelvű statisztikai alapú gépi beszédfelismerés esetén. A koartikulációs modellek integrálhatóságának érdekében súlyozott véges állapotú átalakító (WFST)-alapú beszédfelismerési hálózatépítést használtunk. Az általunk elérhető legnagyobb magyar nyelvű – részben publikus – telefonbeszéd-adatbázisok segítségével értékeltük ki a koartikulációs modelleket, ahol referenciaként általában az implicit koartikulációs modelleket alkalmaztunk.

Megállapítottuk, hogy a *fonológiai* koartikuláció explicit modellezése következetes tanítóadatbázis-feldolgozás mellett nem jár előnnyel az implicit modellezéssel szemben. Ugyanakkor a *fonetikai* koartikuláció explicit modellezésénél mért felismerési eredmények már igen jelentősen túlhaladják az implicit (monofón) modellekkel kapottakat. Különösen a tanításhoz nem illeszkedő teszthalmaz esetén láthatunk drasztikus változásokat az explicit (trifón) modell hatására: itt a *felismerési arányok* nőttek több mint kétszeresükre. Összességében tehát arra jutottunk, hogy a magyar nyelvű statisztikai gépi beszédfelismerésben a fonológiainak a fonetikai koartikulációtól való megkülönböztetésére nincs szükség, míg „a koartikuláció" explicit modellezése úgymond „létszükséglet".

Fontos megjegyezni, hogy az ML döntési fa-alapú trifón állapotcsoportosításnak köszönhetően nem szükséges több száz óra tanítóadatbázis a környezetfüggő beszédhangmodellek betanításához. Amint az eredmények mutatják, a legkisebb, alig pár órás tanítóadatbázis mellett is nagyon jó felismerési eredmények érhetők el. Ezt az adatbázis gondos tervezésének és kifinomult kialakításának tulajdonítjuk, melyet ezúton is szeretnénk megköszönni az alkotóinak.

Bibliográfia

1. Church, K. W. – Gale, W. A. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 5 (1991) 19–54.
2. Mihajlik, P – Tobler, Z. – Tüske, Z. – Gordos, G. "Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech" in In Proc. of InterSpeech'05, Vol 1, pp. 2677-2680, Lisbon, September (2005)
3. Mihajlik, P – Tatai, P. – Gordos, G. "Automatic Phonetic Transcription and Its Application in Speech Recogniser Training – A case study for Hungarian", IOS Press, Amsterdam, NATO ASI series, under the title "Dynamics of speech production and perception;" co-edited by Georg Meyer and Pierre Divenyi, (2006)